Estimation of Traffic Flow Characteristics from Sampled Data

Lili Yang

Department of Statistics

University of Michigan

QPRC, June 4 2007

Joint with Professor George Michailidis

Outline

- Introduction
- General Framework and Nonparametric Model
- Adaptive EM Algorithm
- Flow Size Estimation
- Mixture Model and Two-Stage EM Algorithm
- Fast Semi-parametric Method

Part I Background

Need for Network Measurements

Match available network resources to demands.

- Evaluate the state of the network;
- Characterize the performance experienced by users;
- Control actions required.

Motivation for Sampling in Network

Problem with the original design of network protocols

- Basic TCP/IP protocol: Best Effort Service Model$\longrightarrow$ highly aggregated data
- Time Sensitive Services (for example: Internet Telephony)$\longrightarrow$ more fined grained measurements: fine time scale, at the traffic flow level.

Solution: Sampling Techniques

- Packet Monitoring: copying a stream of packets from the internal fast, then selecting, storing, analyzing and exporting information on these packets.
- Flow Monitoring: collect statistics at flow level; heavy-tailed nature (Willinger,W. (1997)):CHALLENGE

Network Sampling Implementation Issues

- Routers can sample packets, NOT flows
- Originally, only systematic sampling available (1/100 packets)
- More recently, probabilistic sampling possible

Part II Estimating Characteristics of Traffic Flows

Understanding the characteristics of traffic flows is crucial for allocating the necessary resources (bandwidth) to accommodate users demand.

Problem Formulation

Suppose on a network link there are M active flows, comprised of $N_m,\ m=1,...,M$ packets each. The number of packets in each flow is referred to as the *flow length*. The payload of each packet consists of $Z_m^{(i)},\ i=1,...,N_m$ bytes and the size of the m-th flow in bytes is given by $B_m=\sum_{i=1}^{N_m} Z_m^{(i)}$, which is referred to as the *flow size*.

Bernoulli sampling scheme

: Observed data are sampled flow lengths $n_1, n_2, ..., n_r$, and their corresponding flow sizes $b_1, b_2, ..., b_r$, with $b_k=\sum_{i=1}^{n_k} Z_k^{(i)}$

$\{n_1, n_2, ..., n_r\}$: $\{j,\ g_j\},\ j=1,...,J$

NOTE: an online implementation of such a sampling scheme yields biased samples for long flows.

Objective:

1. estimate non-parametrically and semi-parametrically the flow length distribution F of the link, and in addition estimate the original length of sampled flows $N_i, i = 1, 2, ..., r$;
2. estimate the flow size (expressed in bytes) distribution G and similarly estimate the original flow sizes $B_i, i = 1, 2, ..., r$;
3. estimate the number of active flows M in the link.

$$G(B_0) = \mathsf{P}(\sum_{i=1}^{N} Z^{(i)} = B_0) = \int_N \mathsf{P}[\sum_{i=1}^{N} Z^{(i)} = B_0|N]dF(N) = \int_N Q(B_0|N)dF(N). \tag{1}$$

Nonparametric Estimation of Flow Length Distribution F

Model

$$L(\phi_i, M) = \binom{M}{g_0, g_1, \ldots g_J} \prod_{j \geq 0} \left(\sum_{i \in S_I} \phi_i c_{ij}\right)^{g_j} \quad (2)$$

Notation

ϕ_i: the probability that a flow contains i packets.

c_{ij}: the probability of having j packets sampled, given the true flow length is i.

$g_j,\ j = 0, 1, \ldots, J$: the frequency of sampled flows of length j.

$S_I = \{i(0), i(1), \ldots, i(J)\}$, with $i(j)$ denoting the length of a flow being i packets when j of them have been sampled. In the initial setting, we choose $i(0) = \llcorner \frac{1}{2p} \lrcorner$ and $i(j) = \llcorner j/p \lrcorner$.

Adaptive EM Algorithm

(1) E-step:

Complete set of data: (f_{ij}, g_j).

Frequency of flows of length i and with j packets sampled f_{ij} follows Multinomial$(M = \sum_{i,j} f_{ij}, p_{ij})$.

$$Q(\phi, \phi^{(k)}) = \sum_{i \geq j \geq 0} E_{\phi^{(k)}}(f_{ij}|g_j, j = 1, 2, ..., J)\text{log}(\phi_i c_{ij}).$$

- $j \neq 0,\ E_{\phi^{(k)}}(f_{ij}|g_j, j = 1, 2, ..., J) = g_j p_{i|j},$
- $j = 0$, nuisance parameter $\hat{g_0}^{(k)} = \sum_{j>0} g_j w_j,$

  where $w_j = \frac{\sum_i \phi_i c_{i0} c_{ij}}{\sum_i \phi_i c_{ij}}$.

  Hence, $E_{\phi^{(k)}}(f_{i0}|g_j, j = 1, 2, ..., J) = \hat{g_0}^{(k)} p_{i|0}$

(2) M-step:

$\phi^{(k+1)} =$ arg max$Q(\phi, \phi^{(k)})$, s.t. $\sum_{i \in S_I^{(k)}} \phi_i = 1,$ and $\phi_i(j) \geq 0$ for $i \in S_I^{(k)}$.

$$\phi_i^{(k+1)} = \frac{\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g_0}^{(k)} p_{i|0}}{\sum_{i \in S_I^{(k)}} (\sum_{i \geq j \geq 1} g_j p_{i|j} + \hat{g_0}^{(k)} p_{i|0})},$$

where $p_{i|j}$ is the conditional probability that for a flow of length i given j of its packets have been sampled.

(3) Adjusting Support Step

Given the estimated flow length distribution ϕ, the posterior probability distribution of a flow being of length i given that j of its packets have been sampled

$f(i|j) = \frac{c_{ij}\phi_i}{\sum_{i \in S_I} c_{ij}\phi_i},\ j = 1, 2, ..., J$

For any given sampled flow of length j, we provide an estimator of the original flow length $i(j)$, substituting the support $S_I^{(k)}$.

$$\hat{i}(j) = \mathsf{E}(i(j)) = \sum_{i \in S_I^{(k)}} i f(i|j). \tag{3}$$

Iterate steps (1) - (3) until the convergence criterion is satisfied; i.e.

$||\phi^{(k+1)} - \phi^{(k)}|| < \delta$.

Estimation of Flow Size Distribution

We have already nonparametrically estimated flow length distribution F by ϕ, the next main issue to estimate flow size distribution G is to estimate $Q(B|N)$ according to previous described general framework(1), where

$$Q(B|N) = \mathsf{P}(\sum_{k=1}^{N} Z^{(k)} = B|N).$$

Sample Point of View: $Q(b|j) = \mathsf{P}(\sum_{k=1}^{j} Z^{(k)} = b|j)$.

Regression Model:

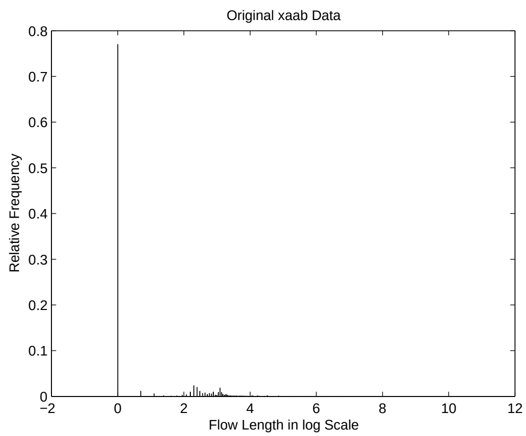
$$b_j = \gamma_0 + \gamma_1 j + \epsilon, \text{ for all } j;$$

$$\hat{B}_j = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{i}(j)$$

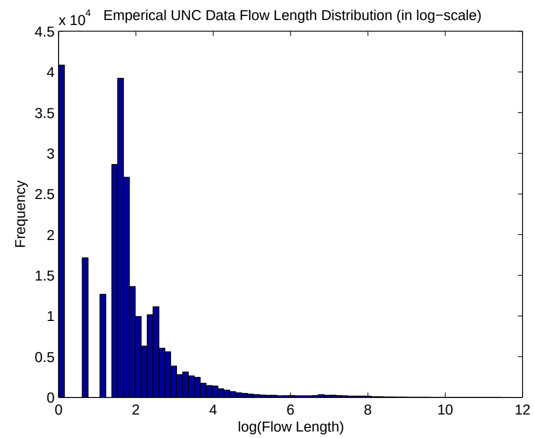
$\hat{B}_j$ gives support of flow size distribution S_B.

Mixture Distributions

Notice that both the packet length and size distribution are mixtures of two components; the first, representing short flows, and the second representing considerably longer flows.



(a) LAN flow length distribution



(b) UNC flow length distribution

Figure 1: Real Data flow lengths distribution in log-scale

Assume the original flow length distribution F (and consequently the flow size one G) is a mixture of two components; i.e.

$$F = \alpha F_1 + (1 - \alpha) F_2,$$

with $\alpha \in (0, 1)$. To keep this simple, further assume $F_1 \equiv \delta_1$.

We propose a Two-Stage EM Algorithm that deals with the problem of estimating mixture distributions.

Two-Stage EM Algorithm

- Adaptive EM Algorithm : estimate ϕ, S_I and M.
- Another EM Algorithm based on the current estimates of these parameters : estimate the mixing coefficient α.

Split the parameters of interest into two subsets (blocks) and in each iteration alternate between the blocks by fixing the parameters of the other block in their current values.

Profile likelihood function for estimating α:

$$\begin{aligned} L(\alpha) &= \binom{M}{g_0, g_1, \dots g_J} \prod_{j \geq 0} (f_j)^{g_j} \\ &\sim \prod_{j \geq 0} [\alpha f(j|1) + (1-\alpha) f(j|S_I^2)]^{g_j}, \end{aligned}$$

where S_I^2 is the support of the second component.

Experimental Evaluation

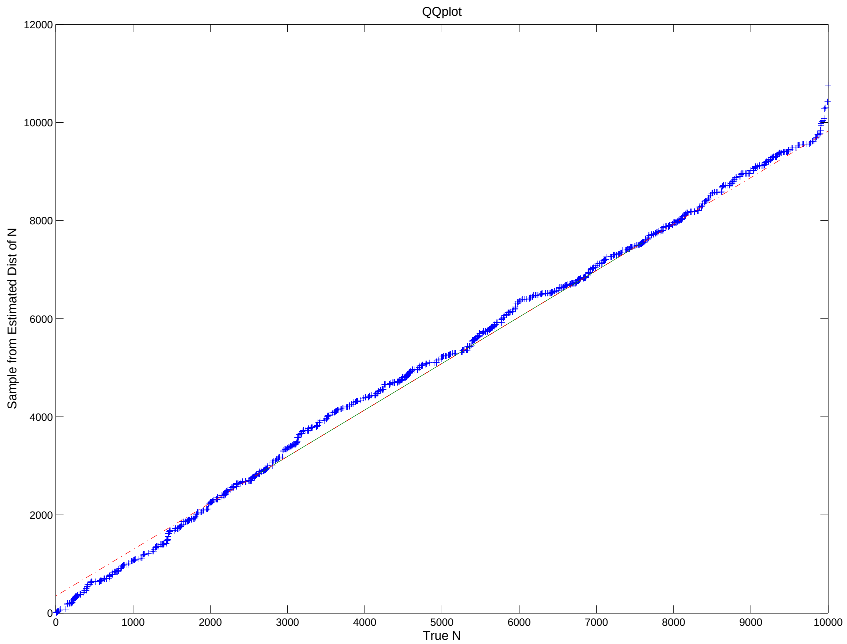


Figure 2: Quantile-quantile plot of the true vs the estimated flow length distribution for 1,000 Uniform flows with .05 sampling rate

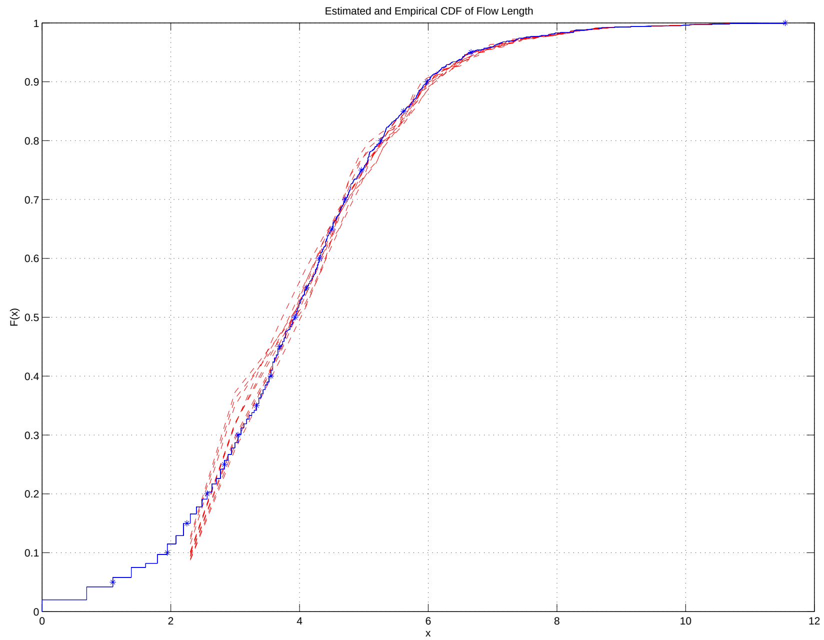


Figure 3: Dash lines are CDF Curves of the estimated flow length distribution for 100 Pareto flows with .05 sampling rate; solid line with '*' is CDF Curve of the true flow length distribution

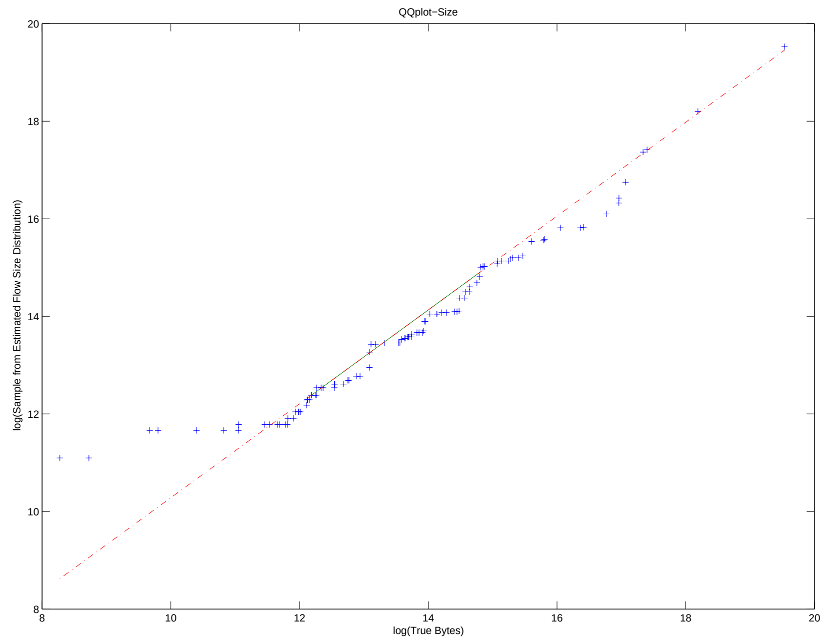


Figure 4: QQplot for true vs estimated flow size distribution from 100 pareto flows with bytes per packet following normal(1350,100); Adaptive EM with sampling rate is p=0.05

The experiment using simulated flow length data from the poisson distributions with mean 5000:

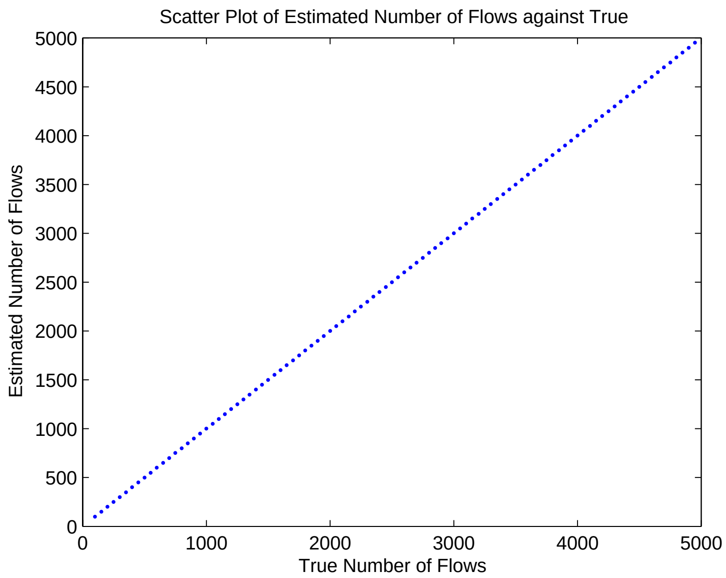


Figure 5: Scatter plot of true vs estimated number of active flows in the link for Poisson flows

A real network trace obtained from the router of the Abilene network at Denver in June of 2005. The trace covers a 5-minute period and contains 65,535 active flows. The average flow length consists of 3 packets, but the variance takes a value of 430.

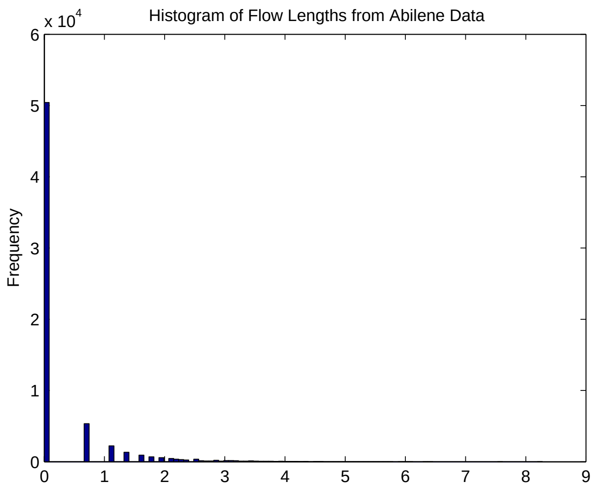


Figure 6: Histograms of Flow Lengths in log scale from NetFlow data

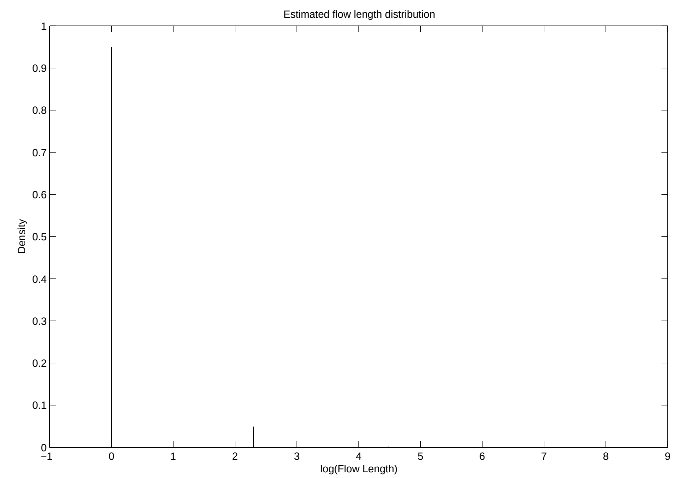


Figure 7: Estimated flow length distribution (in log-scale) of Abilene trace using a 2-stage EM algorithm

Semi-parametric Methods

Slow convergence for large data set $\longrightarrow$ A faster alternative to capture the flow characteristics of the first two moments. Application: Anomaly Detection.

Theoretically sample moments system:

$$\mu_n = p\mu_N,\ \sigma_n^2 = p(1-p)\mu_N + p^2\sigma_N^2,$$
$$\mu_b = \mu_n\mu_Z,\ \sigma_b^2 = \sigma_Z^2\mu_n + \mu_Z^2\sigma_n^2,$$
$$\text{cov}(b,n) = \mu_Z\sigma_n^2.$$

(μ_N, σ_N^2): the first two moments of the number of packets on the mth flow;

(μ_n, σ_n^2): the first two moments of the number of packets sampled on the mth flow;

(μ_z, σ_z^2): the first two moments of bytes/packet on the mth flow;

(μ_B, σ_B^2): the first two moments of total bytes on the mth flow.

(μ_b, σ_b^2): the first two moments of total bytes observed on the mth flow.

Method of Moment (MM)

Moment of flow lengths: $\hat{\mu}_N^{(MM)} = \hat{\mu}_n/p$, $\hat{\sigma}_N^{2(MM)} = \frac{\hat{\sigma}_n^2 - (1-p)\hat{\mu}_n}{p^2}$

Moment of flow size:

$$\hat{\mu}_B^{(MM)} = \hat{\mu}_Z^{(MM)} \hat{\mu}_N^{(MM)},$$

$$\hat{\sigma}_B^{2(MM)} = (\hat{\sigma}_Z^{2(MM)} + \mu_Z^2)(\hat{\sigma}_N^{2(MM)} + \mu_N^2) - (\mu_Z \mu_N)^2,$$

where the independence between N and Z is assumed.

Moment Least Square (MLS)

The estimates are achieved by minimizing

$$\begin{aligned} L(\mu_N, \sigma_N^2, \mu_Z, \sigma_Z^2) \quad = \quad & [\hat{\mu}_n - p\mu_N]^2 + [\hat{\sigma}_n^2 - p(1-p)\mu_N + p^2\sigma_N^2]^2 + [\hat{\mu}_b - \hat{\mu}_n \mu_Z]^2 \\ & + [\hat{\sigma}_b^2 - \sigma_Z^2 \hat{\mu}_n + \mu_Z^2 \hat{\sigma}_n^2]^2 + [\widehat{\text{cov}}(b, n) - \mu_Z \hat{\sigma}_n^2]^2 \end{aligned}$$

Bias Correction

Bias:

- $\mathsf{E}(n) = p\mu_N$ is estimated by $\hat{\mu}_n = \sum_m n_m/r$, an unbiased estimator of $\mathsf{E}(n|n>0) = p\mathsf{E}_N(\frac{N}{1-(1-p)^N})$.
- Similarly, $\sigma_n^2 = p(1-p)\mu_N + p^2\sigma_N^2$ is estimated by sample variance of all positive sample lengths flows, which is essentially unbiased estimate of var$(n|n>0)$.

Solution:

- Estimating the total number of active flows M by $\hat{M} = \frac{r}{1-c_{\hat{\mu}_N 0}}$, where $\hat{\mu}_N$ is from the estimated average flow lengths without bias-correction.
- Next, $\hat{\mu}_n$ is updated by $\sum_m n_m/\hat{M}$ to accommodate the unobserved flows.
- Subsequently, MM or MLS can be applied with new $\hat{\mu}_n$.

This gives a more robust estimate than original MM or MLS methods.

Table 1: Empirical Result based on Lognormal Flow Length Distribution

	mean(mN)	mean(var(N))	CI(mN)	mean(mN)	mean(var(N))	CI(mN)
Real	500	50000		5000	5000000	
Method of Moment	513.47	44870	46	5009	5.01E+06	95
Bias Correction	510.57	46095	59	5009	5.01E+06	95
	mean(mean(B))	mean(vB)	CI(mB)	mean(mean(B))	mean(vB)	CI(mB)
Real	3.79E+05	6.91E+10		3.83E+06	7.25E+12	
Method of Moment	3.84E+05	6.46E+10	84	3.75E+06	6.65E+12	81
Bias Correction	3.82E+05	6.51E+10	85	3.75E+06	6.65E+12	81

Conclusions

- The previous work was motivated by the problem of estimating the flow length and size distributions from sampled data.
- A maximum likelihood non-parametric estimator for these quantities is proposed based on Bernoulli sampling and their properties briefly discussed.
- Mixture distributions are considered which are prevalent in real network traffic traces. A two-stage maximum likelihood estimator is proposed based on Bernoulli sampling.
- Fast Semi-parametric methods are discussed to accommodate online anomaly detection.
- Experimental evidence suggests that the quality of the estimates is very good and obviously improves for larger sampling rates.